基于图像视野划分的公共场所人群计数模型 *

袁 健，王姗姗，罗英伟

(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要：为解决公共场所中人群分布不均以及目标尺度不一而影响人数估计的问题，提出了基于图像视野划分的公共场所人群计数模型。该模型首先将图像场景划分为远近视野两个区域：对近视野区域，使用基于 YOLO 的网络进行行人检测并通过添加场景约束避免在远近视野区域内重复计数；对远视野区域，使用改进的 MobileNets 提取人群密度分布特征，并引入超分辨率重建模块提升人群密度图质量，最终通过计算两者之和得到整幅图像中的人群数量。在 Shanghai Tech 和 Mall 数据集上进行测试，结果表明该模型在准确性和鲁棒性上有显著的提高。实验证明，模型切实可行。

关键词：人数估计；卷积神经网络；轻量型

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2020.02.0076

Public place crowd counting model based on image field division

Yuan Jian, Wang Shanshan, Luo Yingwei

(*School of Optical Electrical & Computer Engineering*, *University of Shanghai for Science& Technology*, *Shanghai* 200093, *China*)

Abstract: In order to solve the problems of uneven population distribution and different target scales affecting the crowd counting in public places, this paper proposed a novel crowd counting model based on image field division. Firstly, it divided the image scene into two parts: the near and far field of vision area. For the near field of vision area, it used the YOLO based network for pedestrian detection and added scene constraints to avoid repeated counting in the near and far field of vision. For the far field of vision area, it used the improved MobileNets to extract the population density distribution characteristics, and introduced the super-resolution reconstruction module to improve the quality of the population density map. Finally, it obtained the population in the whole image by calculating the sum of the two. This paper tested the proposed model on Shanghai Tech and Mall datasets, and the results show that the model has a significant improvement in accuracy and robustness. Experiments show that the model is feasible.

Key words: crowd counting; convolutional neural network; lightweight

0 引言

随着社会经济发展以及人口数量的不断增加，人群活动呈现多样性，大型社会聚集活动日益增多，车站、景点、商场等公共场所人群聚集拥挤的场景随处可见，这给公共管理以及公共安全带来巨大的挑战。人群密度往往与人群安全密切相关，一旦某处人群密度过高，拥挤的人群容易引起恐慌，甚至引发踩踏事件，而传统的视频监控系统需要专人守候检测，耗费大量的人力，若能让计算机实时对当前场景中的聚集人群的数量进行监测分析，一旦出现拥挤趋势就自动发出警报，及时通知相关部门进行干预，这对保证公共场所人群安全具有重大意义。但是公共场所情况复杂且环境不可控、人群分布无规律、相互遮挡、光照不均匀、相机透视等问题的存在，导致准确估计人群数量仍然是一项具有挑战性的工作。本文对公共场所人群计数问题进行研究，提出了一种精确度更高，计算速度更快的对图像进行人数识别的公共场所人群计数模型。

1 相关工作

现如今越来越多研究人员开始关注人群计数问题，目前对于此类问题的研究大致可以分为基于行人检测、基于回归以及基于深度学习的三类方法。

早期的人群研究通常以提取整体或局部特征的方式进行。文献[2，3，5~7]通过提取行人全身的特征(如 Haar 小波、HOG、边缘特征)训练分类器进行检测，该类算法在人数不多、密度比较低的人群计数中能够呈现较好的效果。文献[4，9，11，12]通过基于局部特征的方法来解决，其中文献[12]通过 Haar 小波变换提取头部轮廓的特征区域，并利用透视变换技术更准确地估计人群大小。这种方法对存在一定遮挡的人群，检测效果有一些提升，但是随着人群密度的升高，人与人之间的遮挡逐渐变得更加严重时，此算法变得更加耗时，而且计算准确度也不够理想。

基于回归的方法[13,14,17~19]通过学习一种低层次特征(如边缘特征[14,18]、纹理特征[18]等)到人群数量的映射，建立图像特征和图像人数的回归模型，从而得到预测的人数。此类方法把人群看做一个整体，成功解决了相互遮挡等问题，但是却忽略了行人空间分布情况的重要性。于是有研究人员[15,16,20]想到将空间信息融入到学习过程中，通过学习图像特征与对应对象密度图之间的线性映射，建立回归模型。文献[15]基于图像 SIFT 特性，采用线性回归的方式得到人群密度的分布图，之后对密度图进行积分计算，最终得到人群数量，这种方法特点是在环境不复杂的条件下，检测速度很快，避免了学习检测和定位单个对象实例的困难。尽管基于密度回归的方式在一定程度上提高了计数精确度，但其本质仍然是通过手工提取人群特征。

近年来，随着科技的发展以及深度学习技术在计算机视

收稿日期：2020-02-14；修回日期：2020-04-03 基金项目：国家自然科学基金资助项目(61775139)

作者简介：袁健(1971-)，女，四川泸州人，副教授，博士，主要研究方向为数据挖掘、网络安全、图像处理、智能交通、隐私保护等；王姗姗(1993-)，女，山东泰安人，硕士研究生，主要研究方向为深度学习、图像处理(sswang2018@163.com)；罗英伟(2000-)，男，湖北松滋人，本科生.

觉领域的广泛应用，大量的基于深度学习的算法被提出。深度卷积神经网络(convolutional neural network, CNN)凭借所表现出来的优异的特征学习能力成为最成功的深度模型之一。研究人员逐渐开始考虑使用以深度卷积网络为代表的深度学习算法来解决复杂场景下的人群计数问题[21~26]。文献[22]提出交替优化密度图估计和人数估计的算法(CrowdCNN)，首次将深度卷积网络应用于跨场景的人群密度估计和人群计数问题。文献[24]提出一种基于空洞卷积神经网络的单列计数模型，该模型在大幅削减网络参数量和网络训练难度的同时，显著提高了人群计数的精度和人群分布密度图的还原度。文献[26]采用类似 inception 架构的模块提取多尺度的人头信息，在每个卷积层都同时使用不同大小的卷积核，最后通过反卷积得到最终的密度图。

综上所述，基于 CNN 的方法大大简化了前景分割、目标检测定位等复杂的工作，但是，对于公共场所中的人群数量估计，上述方法仍然存在一些不足:

a) 基于 CNN 的算法本质上还是基于回归的方式，这类方法更适用于人群密度分布相对均匀的场景。但是现实生活中公共场所中人群流向具有较大的随机性，往往会呈现高密度和低密度共存的特点，并且公共场所中摄像头通常被放置在高于人群的地方，加之不同的拍摄角度，所获得的人群图像存在各种各样的视角，分布在图像视野不同区域也会有不同的尺度变化，因此对于以上所提到的这种非统一的场景，上述算法并不具有普适性。

b) 现有基于 CNN 的算法通常通过设置多列大小不同卷积核的网络来解决计数过程中的人群尺度变化，相互遮挡等问题，这种做法却导致网络变宽变深，而且在训练过程中还需要不断调整卷积核大小以适应人群尺度变化，因此使网络计算量过大，场景适应性变差，无法进行实时的人群计数预测。

基于以上分析，本文提出一种基于图像视野划分的公共场所人群计数模型(public place crowd counting model based on image field division，简称 IFDM 模型)。该模型以更强的场景适应性，较高精度的计算能力，更小的网络规模，实现对公共场所人群数量的准确估计。经实验验证，模型拥有较好的泛化能力和较强的鲁棒性。

2　IFDM 模型总体结构

IFDM 模型总体结构如图 1 所示，首先对人群图像根据其深度信息图进行远近视野区域划分，图像的深度信息包含了物体相对前后位置信息，能够反映物体距离拍摄源的远近。IFDM 模型使用文献[27]中的方法获取单张图像的深度信息，然后由深度信息颜色的局部相似度[28]，根据局部像素聚类边界将图像划分为远近视野两个区域。对近视野区域，提出了使用添加场景约束的行人检测计数算法(Counting algorithm of pedestrian detection with scene constraints,简称 SCPD)，通过基于 YOLO 的网络进行行人检测并通过添加场景约束避免模型在远近视野区域内重复计数；对远视野区域，提出了一种高质量密度图回归积分计数算法(Regression integral counting algorithm with high quality density map,简称 HQDPRI)，通过设计了一种结合超分辨率重建模块的轻量型网络提取人群密度分布特征并通过映射生成高质量人群密度图，最终通过计算远近视野图像中人数之和得到整幅图像中的人群数量。

3　添加场景约束的行人检测计数算法 SCPD

当根据深度图像聚类的边界提取出图像的分割线后，将其映射到原始人群图像中进行区域的分割，区域划分结果如图 2 所示。由图 2(d)可知，近视野区域行人个体特征明显，信息丰富，人群遮挡不算严重，本文首先用传统卷积网络进行常规的特征学习训练，卷积神经网络以静态人群图像为输入，预先训练的权值生成视觉特征之后就采用 YOLO 架构[29]作为检测模块。但是在实验过程中发现，在根据深度信息进行图像分割的时候，切割线往往会将分割线附近的人分割成两半，模型在近视野区域和远视野区域内可能会出现重复计数情况，为解决这个问题，本文基于 YOLO 网络[29]的行人检测算法提出了添加空间约束的近视野计数算法 SCPD,该算法在 YOLO 网络进行检测后，将中心坐标落在限制范围之内(即无效区域)的检测框删掉，从而避免重复计数，降低错误检测。

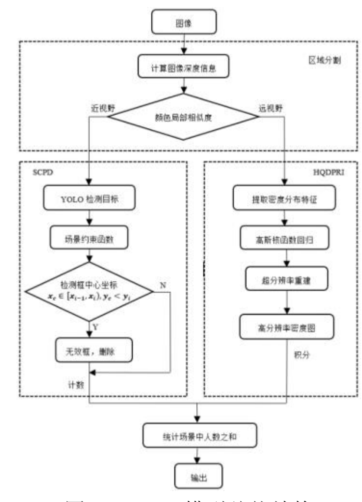


图 1　IFDM 模型总体结构

Fig. 1　Overall structure of IFDM model



(a)原始图像　　(b)深度信息图





(c)远视野区域　　(d)近视野区域

图 2　视野区域示

Fig. 2　Schematic diagram of field of view

首先 YOLO 网络将输入图像划分为 S*S 个单元格，每个单元格给定 B 个不同规格的初始候选框，预测候选框由卷积神经网络提取出来，每幅图像的候选框数量为 S*S*B，同时将预测候选框中是否存在待判别目标的置信度设为

$$Conf(object) = \Pr(object) * \Pr(Class_i \mid Object) * IOU_{pred}^{truth} \tag{1}$$

其中，$\Pr(object)$ 判断网格中是否有需要检测的目标，$\Pr(Class_i \mid Object)$ 表示一个候选框在包含目标的条件下，目标类别为 $Class_i$ 的概率，IOU_{pred}^{truth} 表示真实框与预测框的交并比。由于大部分候选框中并不包含行人，甚至是不包含任何目标物，因此为了减轻网络学习的难度，将不存在目标物的候选框置信度 $Conf(Object)$ 设置为 0，同时由于只需要对存在目标物的候选框进行行人的判别，将目标类别 $Class_i$ 设置为 1。SCPD 算法具体步骤如下：

a) 输入待检测的人群图像。

b) 根据人群空间分布的先验信息，使用式(2)对可能出现误检的区域进行划分。

$$\begin{aligned} y_1 &= k_1 + b_1 \quad x \in [x_0, x_1) \\ y_2 &= k_2 x + b_2 \quad x \in [x_1, x_2] \\ &\vdots \\ y_n &= k_n x + b_n \quad x \in [x_{n-1}, x_n] \end{aligned} \tag{2}$$

折线方程随着场景的不同而变化，设定在切割线与折线之内的区域为无效区域。将目标为行人的概率公式设为 Pr(person|object)，则候选框中包含行人的置信度 Conf (person) 表示为

$$\text{Conf}(\text{person}) = \Pr(\text{object}) \times \Pr(\text{person}|\text{object}) \times \text{IOU}_{\text{pred}}^{\text{truth}} \tag{3}$$

对检测结果进行后续处理。假设检测框的左上角坐标和右下角坐标分别为 $(x_{\min}, y_{\min})$，$(x_{\max}, y_{\max})$，那么这个边界框的中心坐标为 $x_c = (x_{\min} + x_{\max})/2$，$y_c = (y_{\min} + y_{\max})/2$。如果 $x_c \in [x_{i-1}, x_i)$，$y_c < y_i$，$i \in (1, n]$，则表示检测框出现在无效区域，那么直接删除无效区域出现的边框，同时不将此人计入最终统计人数中，添加约束前后检测结果对比如图 3 所示。

c) 对添加约束后的检测结果进行统计输出近视野区域相应的人数。




图 3　添加约束前后检测结果对比

Fig. 3　Comparison of test results before and after adding constraints

4　高质量密度图回归积分计数算法 HQDPRI

不同密度的人群在特征上存在较为明显的差异，远视野区域人群往往尺度较小，相互遮挡比较严重，目标检测的方式在此部分的检测效果并不理想，因此该部分采用密度图回归的方式进行计算。文献[21]通过一种多列网络并联的网络模型实现提取不同尺度的人头特征，但是却导致参数量过大，并且产生了很多低效的分支结构。文献[20]先将图像分块，然后将每个块通过分类网络决定进一步输入到哪个子网络，虽然取得了不错的检测效果，但是却存在与文献[21]同样的问题，不但计算量大，而且简单的分块也影响了计数预测的准确性。因此，本文提出 HQDPRI 算法，通过一种结合超分辨率重建模块的轻量型网络提取人群密度分布特征并通过映射生成高质量人群密度图，最后对高质量密度图进行积分来求出此部分的人数。

4.1　结合超分辨率重建模块的轻量型深度卷积网络

虽然区域划分工作使得本部分不再需要考虑不断调整卷积核大小以适应人群尺度变化，但是远视野区域仍然存在着人群分布密集、相互遮挡等问题。HQDPRI 在改进的轻量级网络 MobileNets[30]提取特征的基础之上，引入了一个超分辨率重建模块，设计了一个新的用于图像人群计数卷积神经网络。

主体网络在 MobileNets 基础上进行改进，以深度卷积和 1×1 的逐点卷积代替标准卷积操作，共设置了 27 层卷积层，由多个 3×3 和 1×1 的卷积核构成。同时减少了步长为 2 的卷积核的个数，将 Conv4 dw 和 Conv5 dw 的步长设置为 1，其余卷积层步长保持不变，这样做的目的是使卷积后的图像尺度更大，保留更多空间细节信息，输入图像大小为 224×224×3，同时去掉了均值池化层和全连接层，最终输出 1/16 原图的密度特征图，具体参数变化如表 1 所示。网络没有采用池化层，而是通过将深度卷积的步长设置为 2 以此实现下采样的目的，这样的组合方式使网络在损失精度不多的情况下大幅度降低了参数量和计算量，提升了检测速度。与常用的 VGG16 网络模型相比，计算准确度与其相似，但是计算复杂度却减小了 27 倍。为了能够获得更加准确的计算精准度，网络后半部分引入一个超分辨率重建模块用于提高密度图的质量。

表 1　主体网络参数表

Tab. 1　Backbone network parameters

卷积层/步长	卷积核	输入尺寸
conv0/s2	3×3×3×32	224×224×3
conv1 dw/s1	3×3×32 dw	112×112×32
conv1/s1	1×1×32×64	112×112×32
conv2 dw/s1	3×3×64 dw	112×112×64
conv2/s1	1×1×32×64	112×112×64
conv3 dw/s2	3×3×64 dw	112×112×64
conv3/s1	1×1×64×128	56×56×64
conv4 dw/s1	3×3×128 dw	56×56×128
conv4/s1	1×1×128×256	56×56×128
conv5 dw/s1	3×3×256 dw	56×56×256
conv5/s1	1×1×256×256	56×56×256
conv6 dw/s2	3×3×256 dw	56×56×256
conv6/s1	1×1×256×512	28×28×256
5×conv dw/s1	3×3×512 dw	28×28×512
5×conv/s1	1×1×512×512	28×28×512
conv12 dw/s2	3×3×512 dw	28×28×512
conv12/s1	1×1×512×1024	28×28×512
conv13 dw/s2	3×3×1024 dw	28×28×1024
conv13/s1	1×1×1024×1024	14×14×1024

4.2　超分辨率重建模块

超分辨重建技术可以实现目标物的专注分析，从而获取感兴趣区域更高空间分辨率的图像，当前基于深度学习的单张图片超分辨率重建在重建效率和计算量方面已经取得了很大的成功。文献[31]提出将低分辨率的图片直接通过卷积网络来做超分辨率，同时提出了一种有效的子像素卷积层，通过学习到一组扩大滤波器去将低分辨率的特征映射到高分辨率的输出。通过这种方式，不但省去了双三次插值法，也大大减轻了计算量。本文在文献[31]基础上进行改进，将超分辨重建技术引入网络结构中，旨在优化密度图质量，从而获得更加准确的计算精准度。

超分辨率重建模块网络的第一层选择使用两个 3×3 的卷积核代替 5×5 的卷积核，这样不仅能够在保证具有同样感知野的条件下提升网络的深度，增加非线性特性的表达，而且在一定程度上也提升了神经网络的特征学习效果。第二层及第三层使用深度可分离卷积代替普通卷积，同时为了适应图像重建任务，省去了 Batch Norm。通过前两层卷积得到特征通道数为 r^2 (r 为图像的目标放大倍数)的与输入图像大小一样的特征图像，随后第三层亚像素卷积层将特征图像的每个像素的 r^2 个通道重新排列成一个 $r \times r$ 的区域，对应于高分辨率图像中一个 $r \times r$ 大小的子块，从而大小为 $H \times W \times r^2$ 的特征图像被重新排列成 $rH \times rW \times 1$ 的高分辨率图像，由此得到优化的人群密度图，该过程实际上并不涉及卷积操作，只是对图像大小做变换，因此效率更高。

4.3　HQDPRI 算法步骤

HQDPRI 算法步骤如下：

a) 将人群图像合集送入改进后的 MobileNets 主体网络中提取卷积特征。

使用带标准差的高斯核函数 $G_{\sigma_i}(x)$ 与头部坐标 $\delta(x - x_i)$ 进行卷积代入式(4)得到人群密度函数 $F(x)$。计算公式为

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x) \quad \sigma_i = \beta \overline{d^i} \tag{4}$$

$$d^i = \frac{1}{m} \sum_{j=1}^{m} d_j^i \tag{5}$$

其中，$\delta(x-x_i)$ 表示坐标为 x_i 的人头标记点，图像中标记点 x_i 的 k 近邻距离分别表示为 $d_1^i, d_2^i, \cdots, d_m^i$，$\overline{d^i}$ 为图像中标记点与其最近的 k 个人头之间的平均距离。实验[21]证明，$\beta=0.3$ 时得到的人群密度图效果最好。

超分辨率重建模块通过亚像素卷积层将密度图 $F(x)$ 的像素重新排列，提高密度图的质量，计算公式为

$$PS(F(x))_{x,y,c} = F(x)_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c\cdot r\cdot \mathrm{mod}(y,r)+c\cdot \mathrm{mod}(x,r)} \tag{6}$$

其中，PS 为一个周期混排算子，它把 $H\times W\times r^2$ 张量的元素后移到形状 $rH\times rW\times 1$ 的张量。$\mathrm{mod}(x,r)$，$\mathrm{mod}(y,r)$ 表示在滤波器卷积期间周期性地激活不同子像素位置上图案，x、y 是高分辨率空间中的输出像素坐标。

b) 通过对高质量密度图进行积分求和来求出此部分的人数 N_p，计算公式为

$$N_p = REG(\sum_{i=1}^{N} x_i) \tag{7}$$

5 实验与分析

为了检验 IFDM 的有效性，选用 Shanghai Tech[21]和 Mall[6]数据集作为实验数据来源，其中 Mall 数据集中的数据信息来自一段商场的视频监控，场景变化较小，人数相对稀少。而 Shanghai Tech 数据集是从网络中随机选择的，人群密度大且场景变化更加丰富。训练集与测试集划分详情如表 2 所示。实验环境基于 Linux 64 Ubuntu16.04 操作系统，深度学习框架使用 TensorFlow，显卡为 GTX-Titan X。

表 2 数据集划分详情

Tab. 2 Details of data set division /幅

数据集	训练集	测试集
Mall	800	1200
ShanghaiTech part_A	300	182
ShanghaiTech part_B	400	316

5.1 模型训练

本文使用欧式距离作为损失函数来测量预测人群密度图与真实密度图之间的差值，计算公式为

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\| F(X_i,\theta) - F_i \right\|^2 \tag{8}$$

其中，θ 为模型训练得到的参数，N 为训练集图片总数，X_i 为输入的第 i 张图像，$F(X_i,\theta)$ 和 F_i 分别代表第 i 张预测人群密度图和真实人群密度图。

为了加快模型收敛速度，本文使用自适应学习率的 Adam 优化算法对网络进行优化，并将初始学习率设置为 1e-5，设 batch_size=4。根据以往的经验，训练集中数据过少在训练过程中容易导致网络过拟合，因为为了避免过拟合现象的产生，本文对训练集中的图片进行处理，即将每张图片裁剪为四个大小相同且互不重叠的块，经过这样的处理之后将训练集扩大了 4 倍。

5.2 评价标准

模型性能使用平均绝对误差(mean absolute error，MAE)和平均平方误差(mean squared error，MSE)来衡量，如式（9）（10）所示。

1)平均绝对误差 MAE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z_i'| \tag{9}$$

其中，N 表示测试集中图片数量，z_i 表示通过预测人群密度图得到的人群数量，z_i' 表示图片中实际的人数。MAE 表示网络预测结果的准确性，MAE 值越小说明估计人群数量越准确。

2)平均平方误差 MSE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z'|| \tag{10}$$

MSE 反映了估计量与被估计量之间的差异程度，MSE 值越小说明算法的鲁棒性越好。

5.3 对比实验结果分析

Shanghai Tech 数据集由 part_A 和 part_B 两部分组成，这两个子集之间存在着显著的密度差异，part_A 中包含 482 张随机从互联网上抓取的图片，人群密度较大; part_B 中包含 716 张从上海繁华街头拍摄的图片，人群密度中等但是人群分布变化较大。此数据集总共包含 1198 张带标注的图片，其中总标记的人头数达到 330165 人。表 3 给出了 Shanghai Tech 数据集上的实验对比结果，文献[20~24]都是基于 CNN 的方法。由表中的结果可以看出，在 part_B 测试集中，该模型的 MAE 与文献[22]的相比下降了 35.94%，MSE 下降了 34.53%；part_A 测试集中的 MAE 下降了 39.38%，MSE 下降了 38.06%。与文献[21]相比，Part_A 中 MAE 与其相持平，与文献[24]相比，part_A 中 MAE 反而增加了 8.52%，这是因为 part_A 测试集中的人群密度较大，难以区分出明显的远近视野区域，但是较其他实验结果可以看出 MSE 仍表现得较为优秀。通过以上数据可以看出，与同样是基于 CNN 的算法进行比较，该模型整体结果优于经典算法，在人群密度分布变化较大的情况下更具有良好的性能。图 4 给出了模型在 Shanghai Tech 数据集的实验结果示意图。

表 3 Shanghai Tech 数据集的实验结果对比

Tab. 3 Comparison of experimental results in Shanghai tech dataset

方法	part_A		part_B	
	MAE	MSE	MAE	MSE
CrowdCNN [22]	181.8	277.7	32.0	49.8
FCN[23]	126.5	173.5	23.7	33.1
MCNN[21]	110.2	173.2	26.4	41.3
HCNN[25]	100.8	152.3	21.5	33.4
IFDM	110.2	172.0	20.5	32.6

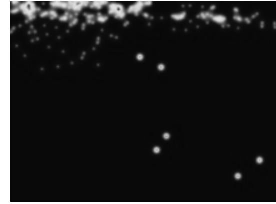



图 4 测试图片密度

Fig. 4 Schematic diagram of test image density

Mall 数据集由 2000 帧大小为 640×480 的帧组成，其中总标记的行人数量超过 60000 人，除了具有不同的光照条件和人群密度之外，数据集的透视畸变较为严重，物体尺寸和外观变化较大，遮挡也更为频繁。表 4 给出了 Mall 数据集上的实验结果，文献[13，17]是基于传统的方法，文献[20，32]是基于 CNN 的方法，由表中的结果可以看出，与传统方法[17]相比，IFDM 模型的 MAE 下降了 40.00%，而 MSE 提升更为明显；与基于 CNN 的方法[20]相比，模型的 MAE 下降了 16.73%，MSE 同样有较明显的改善。Mall 数据集场景变化相对固定，单幅图像中的人数相对稀少，实验结果表明模型在人群相对稀疏的图像进行估计也能获得较精确的结果，而且具有更高的鲁棒性。图 5 给出了模型在 Mall 数据集的实验结果示意图。

表 4 Mall 数据集的实验结果对比

Tab. 4 Comparison of experimental results in the mall dataset

方法	MAE	MSE
CARR[17]	3.43	17.7
GPR[13]	3.72	20.1
MoC-CNN [32]	2.75	13.4
VLAD-CNN[20]	2.86	13.1
IFDM	2.45	3.2




图 5　测试图片密度

Fig. 5　Schematic diagram of test image density

5.4　验证性实验分析

为验证超分辨率重建模块对模型性能影响，本节主要对去掉超分辨率重建模块之后模型的运行速度以及性能指标两方面进行验证分析。表 5 给出了本文算法与不添加超分辨率重建模块的算法在 Shanghai Tech 数据集 part_B 上的性能指标对比，由表 5 数据可以看出，对比本文算法，不添加超分辨率重建模块的算法 MAE 降低了 39.5%，MSE 降低了 53.3%。表 6 给出了在输入图像大小为 224×224 的条件下，有无超分辨率重建模块的模型总参数量、总计算量以及模型运行速度对比结果，由表 6 数据可以看出，模型添加超分辨率重建模块后，参数量以及计算量并没有大幅增加，这是因为主体网络作为轻量级网络，本身参数量与计算量比起常规网络就少的多，而且本文对子像素卷积层进行了改进，同样大大减少了参数量与模型的计算复杂度，因此对于本文添加超分辨率重建模块的模型，仍能保持较快的运行速度。

综上所述，引入了超分辨率重建模块的模型，虽然增加了一定的计算量，使模型运行速度较无此模块的有所降低，但是能有效提高预测人群密度图的质量，使模型的性能指标明显增加，能得到更加准确的预测结果。

表 5　有无超分辨率重建模块性能对比

Tab. 5　Performance comparison of super-resolution reconstruction module　/%

方法	MAE(%)	MSE(%)
有超分辨率重建模块	20.5	32.6
无超分辨率重建模块	28.6	50.0

表 6　有无超分辨率重建模块参数量、计算量以及运行速度对比

Tab. 6　Comparison of parameters, calculation and operation speed of super-resolution reconstruction module

方法	参数量/百万	计算量/百万	运行速度/fps
有超分辨率重建模块	0.29	39.96	52
无超分辨率重建模块	0.23	32.27	48

6　结束语

公共场所人群计数问题是人群行为研究中一个具有挑战性的课题，也是公共安全领域的研究重点。公共场所中往往包含多个不同的物体同时移动，这些物体的尺寸通常较小，并且在图像中呈现出类似的外观，同时还存在相互遮挡，光照不均、相机畸变等因素，这些因素使得公共场所人群数量分析变得非常困难。为了更好的解决这一问题，提出针对不同的视野区域采用不同的计数算法，最后通过计算远近视野图像中人数之和得到最终的预测人数。通过实验分析可知，本文提出的模型虽然较现有方法有了较明显的改善和提升，但是对于一些极端密集的场景，尤其是难以划分远近视野区域的场景仍然存在一些问题。本文后续工作准备继续改进网络结构以适应人群极端密集的场景，同时希望能使模型能够应用到实时视频图像的分析中，通过自动可靠地获取监控中的人数或人群密度，对人群的流动状态、流动方向和持续时间作出综合动态预估，帮助工作人员优化管理。

参考文献：

[1] 张君军，石志广，李吉成. 人数统计与人群密度估计技术研究现状与趋势 [J]. 计算机工程与科学, 2018, 40 (2): 282-291. (Zhang Junjun, Shi Zhiguang, Li Jicheng. Current researches and future perspectives of crowd counting and crowd density estimation technology [J]. Computer Engineering and Science, 2018, 40 (2): 282-291.)

[2] Wojek C, Dollar P, Schiele B, *et al.* Pedestrian Detection: An Evaluation of the State of the Art [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34 (4): 743-761.

[3] Enzweiler M, Gavrila D M. Monocular pedestrian detection: survey and experiments [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31: 2179-2195.

[4] Li Min, Zhang Zhaoxiang, Huang Kaiqi, *et al.* Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection [C]// International Conference on Pattern Recognition. IEEE, 2009: 1998-2001.

[5] Leibe B, Seemann E, Schiele B. Pedestrian Detection in Crowded Scenes [C]// Proc of Computer Vision and Pattern Recognition, 2005: 878-885.

[6] Chen Ke, Loy C, Gong Shaogang, *et al.* Feature Mining for Localised Crowd Counting [C]// British Machine Vision Conference, 2012: 3.

[7] 陈锐，彭启民. 基于稳定区域梯度方向直方图的行人检测方法 [J]. 计算机辅助设计与图形学学报, 2012, 24 (3): 372-377. (Chen Rui, Peng Qimin, Pedestrian detection method based on gradient direction histogram of stable region [J]. Journal of Computer-Aided Design and Computer Graphics, 2012, 24 (3): 372-377.)

[8] 樊春年，杜卫平，刘艳荣. 基于 HOG 特征结合 Adaboost 算法的行人检测 [J]. 自动化技术与应用, 2018, 37 (07): 89-91. (Fan Chunnian, Du Weiping, Liu Yanrong. Pedestrian detection based on HOG features and AdaBoost algorithm [J]. Automation technology and application, 2018, 37 (07): 89-91.)

[9] Wu Bo, Nevatia R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors [C]// the 10th IEEE International Conference on Computer Vision. ICCV, 2005: 90-97.

[10] Viola P, Jones M J. Robust Real-Time Face Detection [J]. International Journal of Computer Vision, 2004, 57 (2): 137-154.

[11] Felzenszwalb P F, Girshick R B, Mcallester D, *et al.* Object Detection with Discriminatively Trained Part-Based Models [J]. IEEE Trans on Software Engineering, 2010, 32 (9): 1627-1645.

[12] Lin Shengfu, Chen J, Chao Hungxin. Estimation of number of people in crowded scenes using perspective transformation [J]. IEEE Trans on Systems Man & Cybernetics Part A-Systems and Humans, 2001, 31 (6): 645-654.

[13] Chan A B, Vasconcelos N. Counting People With Low-Level Features and Bayesian Regression [J]. IEEE Trans on Image Processing, 2012, 21 (4): 2160-2177.

[14] Ryan D, Denman S, Fookes C, *et al.* Crowd Counting Using Multiple Local Features [C]// Digital Image Computing: Techniques and Applications. IEEE, 2009: 81-88.

[15] Lempitsky V S, Zisserman A. Learning To Count Objects in Images [C]// Neural Information Processing Systems, 2010: 1324-1332.

[16] Fiaschi L, Nair R, Koethe U, *et al.* Learning to count with regression forest and structured labels [C]// International Conference on Pattern Recognition, 2012: 2685-2688.

[17] Chen Ke, Gong Shaogang, Xiang Tao, *et al.* Cumulative attribute space for age and crowd density estimation [C]// Proc of Computer Vision and Pattern Recognition, 2013: 2467–2474.

[18] 李海丰，姜子政，范龙飞，等. 基于密度分类及组合特征的人数估计算法 [J]. 计算机应用研究, 2018, 35 (06): 1891-1895. (Li Haifeng, Jiang Zizheng, fan Longfei, *et al.* Population estimation algorithm based on density classification and combination characteristics [J]. Computer application research, 2018, 35 (06): 1891-1895)

[19] Li Jun, Tao Dacheng. A Bayesian Hierarchical Factorization Model for Vector Fields [J]. IEEE Trans on Image Processing, 2013, 22 (11): 4510-4521.

[20] Sheng Biyun, Shen Chunhua, Lin Guosheng, *et al.* Crowd counting via weighted VLAD on dense attribute feature maps [J]. IEEE Trans on Circuits and Systems for Video Technology, 2018: 1788-1797.

[21] Zhang Yingying, Zhou Desen, Chen Siqi, *et al.* Single-Image Crowd Counting via Multi-Column Convolutional Neural Network [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2016: 589-597.

[22] Zhang Cong, Li Hongsheng, Wang Xiaogang, *et al.* Cross-scene crowd counting via deep convolutional neural networks [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 833-841.

[23] Marsden M, Mcguinness K, Little S, *et al.* Fully Convolutional Crowd Counting on Highly Congested Scenes [C]// the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) , 2016: 27-33.

[24] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1091-1100.

[25] 范绿源，仝明磊，李敏，等. 静态图像中采用混合卷积结构进行人群密度估计 [J/OL]. 计算机应用研究: 1-6 [2020-03-25]. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 06. 0661. (Fan Lvyuan, Tong Minglei, Li Min, *et al.* Population density estimation using mixed convolution structure in still images [J/OL]. Computer application research: 1-6 [2020-03-25] https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 06. 0661)

[26] Cao Xinkun, Wang Zhipeng, Zhao Yanyun, *et al.* Scale aggregation network for accurate and efficient crowd counting [C]// Computer Vision. 15th European Conference (ECCV 2018) , 2018: 757-763.

[27] Eigen D, Puhrsch C, Fergus R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network [C]// the 28th Conference on Neural Information Processing Systems (NIPS) , 2014, 27: 2366–2374.

[28] Achanta R, Shaji K, Smith, *et al.* SLIC superpixels. EPFL Technical Report 149300, 2010.

[29] Redmon J, Divvala S, Girshick R, *et al.* You only look once: Unified, real-time object detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) . 2015: 779–788.

[30] Howard A G, Zhu Menglong, Chen Bo, *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [C]// Computer Vision and Pattern Recognition. 2017: 1-9.

[31] Shi Wenshi, Caballero J, Huszár F, *et al.* Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2016: 1874-1883.

[32] Kumagai S, Hotta K, Kurita T. Mixture of counting CNNs: adaptive integration of CNNs specialized to specific appearance for crowd counting [C]// Proc of Computer Vision and Pattern Recognition, arXiv: 1703. 09393, 2017.